Adaptive Template Transformer for Mitochondria Segmentation in Electron Microscopy Images

Yuwen Pan[1*] Naisong Luo[1*] Rui Sun[1] Meng Meng [1]
Tianzhu Zhang[1,2†] Zhiwei Xiong[1,2] Yongdong Zhang[1,3]
[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]State Key Laboratory of Communication Content Cognition, People's Daily Online
{panyw, lns6, issunrui, meng18}@mail.ustc.edu.cn {tzzhang, zwxiong, zhyd73}@ustc.edu.cn

Abstract

Mitochondria, as tiny structures within the cell, are of significant importance in studying cell functions for biological and clinical analysis. And exploring how to automatically segment mitochondria in electron microscopy (EM) images has attracted increasing attention. However, most of existing methods struggle to adapt to different scales and appearances of the input due to the inherent limitations of the traditional CNN architecture. To mitigate these limitations, we propose a novel adaptive template transformer (ATFormer) for mitochondria segmentation. The proposed ATFormer model enjoys several merits. First, the designed structural template learning module can acquire appearance-adaptive templates of background, foreground and contour to sense the characteristics of different shapes of mitochondria. And we further adopt an optimal transport algorithm to enlarge the discrepancy among diverse templates to activate corresponding regions fully. Second, we introduce a hierarchical attention learning mechanism to absorb multi-level information for templates to be adaptive scale-aware classifiers for dense prediction. Extensive experimental results on three challenging benchmarks including MitoEM, Lucchi and NucMM-Z datasets demonstrate that our ATFormer performs favorably against state-of-the-art mitochondria segmentation methods.

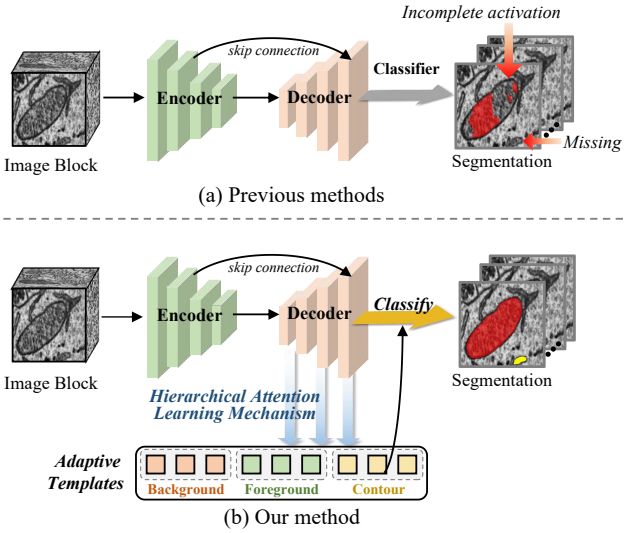


Figure 1. Comparison of the previous methods and our method. (a) Most existing methods use the traditional CNN architecture to extract features, with a simple and weight-fixed classifier at the end of the network for per-voxel classification, resulting in incomplete or missing activation. (b) We propose a set of learnable vectors as adaptive templates of background, foreground and contour, which interact with multi-scale features to aggregate structure-aware information. Then, these templates are used as appearance-adaptive scale-aware classifiers to generate more accurate activation for mitochondria of different scales.

1. Introduction

Studying the morphology of mitochondria is vital for understanding cell physiology, and changes in their shape are tightly linked to neurodegeneration, lifespan and cell death [6, 4, 19, 36]. As representative membrane-bound organelles, mitochondria provide power for cells as the main place to perform aerobic respiration [49, 36, 5, 52], however, their micrometer size makes it challenging to conduct observational analysis. Electron microscopy (EM), as one of the practical tools for investigating the morphology of structures at the sub-cubic millimeter scale [38, 41], enables neuroscientists to obtain images with higher resolution. Due to its large data size and numerous cluttered irrelevant organelles, manual labeling is extremely time-

*Equal contribution
†Corresponding author.

consuming and labor-intensive, which leads to the persistent desire for automatic segmentation algorithms of mitochondria.

With significant advances in deep learning (DL), DL-based methods have paved the way for new research directions toward automated mitochondria segmentation. As a representative work, Liu *et al.* [24] produce segmentation results of mitochondria after the detection process inspired by Mask R-CNN [15], while Cellpose [40] adopts a standard U-Net [37] to generate final predictions. Notably, the above methods only utilize 2D networks to learn the representation of volumetric EM images, neglecting the association between slices in 3D space. To alleviate this issue, recent works design the fully automated pipeline for segmentation based on 3D convolutional network [54, 52], in which several consecutive slices are leveraged as input to achieve promising results.

Recently, some methods in biomedical image analysis [9, 14, 2, 13, 16] introduce the transformer [46] architecture for feature encoding. However, they still use the conventional decoder for upsampling and a convolution-based classifier for prediction, thus fail to explore the full potential of transformers, leaving two issues to be solved: (1) **Unsuitable to handle large appearance variations.** When dealing with diverse input images during inference, it is hard to handle large variations in appearance by adopting a weight-fixed classifier as shown in Figure 1(a). Especially in real EM images, mitochondria exist in a variety of appearances, such as rings, threads, dumbbells, etc. Thus it is essential to empower the classifier to be appearance-adaptive to sense the characteristics of different types of mitochondria. (2) **Unaware of scale variations.** Existing methods leverage a conventional CNN-based classifier for dense prediction, which cannot adapt to large-scale variations and fully exploit the hierarchical information to recognize objects with significantly different sizes in 3D space accurately. Therefore, it is crucial to design scale-friendly classifiers in order to achieve a better performance with more accurate segmentation of mitochondria of all sizes.

In this paper, we propose a novel **A**daptive **T**emplate Trans**former** (**ATFormer**) to obtain adaptive scale-aware classifiers tailored for mitochondria segmentation, including a structural template learning module and a hierarchical attention learning mechanism. **In the structural template learning module,** to obtain *appearance-adaptive* classifiers and capture structure-aware information, we design three groups of learnable templates aiming to three specific categories (background, foreground and contour). Each category corresponds to several complementary templates rather than a single one. Benefiting from the strong ability of transformers, we can obtain appearance-adaptive templates by cross-attention with voxel embeddings. Besides, to avoid the possibility of different templates perceiving repetitive area information, we further propose a regularization term with an optimal transport algorithm to enlarge the discrepancy among diverse templates. **In the hierarchical attention learning mechanism,** in order to attain *scale-aware* classifiers from coarse to fine level, we elegantly design several attention gates which enhance the context extraction by self-attention and transport multi-level features into each layer of the structural template learning module, making templates absorb hierarchical information. In this way, the structural templates then aggregate explicit sufficient semantic information and evolve into adaptive scale-aware classifiers, so that our network can further strengthen the integrity region with finer activation of the whole mitochondria of different scales.

To sum up, our contributions can be summarized as follows:

- We propose a novel adaptive template transformer (ATFormer) for mitochondria segmentation in EM images. Specifically, we design the structural template learning module to acquire adaptive templates of background, foreground and contour for dense predictions, and the hierarchical attention learning mechanism to make the templates adapt to mitochondria of different scales.
- To make the distribution of adaptive templates more discrete and diverse, we further adopt a regularization term with an optimal transport algorithm for full activation.
- Extensive experimental results on three challenging benchmarks including MitoEM, Lucchi and NucMM-Z demonstrate that our proposed method performs favorably against state-of-the-art mitochondria segmentation methods.

2. Related Work

2.1. Mitochondria Segmentation

Mitochondria segmentation is of enormous biological importance for the study of cellular functions and subcellular activities. Rather than early works utilizing traditional image processing techniques [47, 29, 39], recent DL-based methods [34, 8, 32, 17] have shown significant performance improvement on mitochondria segmentation. For example, Oztel *et al.* [34] first segment mitochondria based on a 2D convolutional network and then aggregate results in the axial dimension. U3D-BC [52] utilizes a 3D U-Net [10] architecture with supervisions of foreground and contour and a post-processing step to produce final instance segmentation, while Res-UNet [20] designs a network consisting of anisotropic convolution blocks to boost the segmentation performance. However, these existing methods leverage a simple weight-fixed classifier for all images during inference, leading to incomplete activation of mitochondria with

significant appearance/scale variations. We hereby propose our ATFormer to obtain adaptive scale-aware classifiers for accurate predictions of different types of mitochondria.

2.2. Vision Transformer

Transformers are first introduced in [46] for machine translation. Since transformers have achieved remarkable success in NLP tasks [18, 56], many efforts have been made to introduce transformers to vision tasks including image classification [46, 50, 45, 31] and biomedical image segmentation [14, 26, 25, 42, 51, 43]. SegFormer [55] utilizes a hierarchical transformer encoder for feature extraction and a lightweight MLP-decoder to fuse features for semantic segmentation. UNETR [14] is the first to utilize ViT [12] for feature extraction in biomedical image analysis. Swin UNETR [44] further adopts a Swin transformer as the encoder in a U-shaped network for medical image segmentation. Tailored for mitochondria segmentation, we design three groups of learnable templates based on the transformer architecture to capture specific semantic information and introduce a hierarchical attention learning mechanism to aggregate multi-scale features, in this way, these templates evolve into input-specific classifiers for dense prediction.

3. Method

In this section, we first present the overview of the proposed ATFormer in Sec. 3.1. Then we describe the details of the structural template learning module in Sec. 3.2 and the hierarchical attention learning mechanism in Sec. 3.3. Finally, in Sec. 3.4 the training and inference procedure are discussed.

3.1. Overview

As shown in Figure 2, given a 3D image block $\mathbf{I} \in \mathbb{R}^{H\times W\times D}$, where H, W, and D refer to the height, width and depth of the input, respectively. The bottom-level feature map $\mathbf{X}$ extracted from the backbone encoder is fed into the upsampling module together with skip connections, which outputs hierarchical voxel embeddings $\mathbf{F} = \{\mathbf{f}_l\}_{l=0}^{4}$ to be modified through attention gates. Then the structural template learning module integrates the modified hierarchical features $\widetilde{\mathbf{F}}$ with the help of an OT regularization term to generate diverse adaptive templates, which are used as classifiers to produce background, foreground and contour activation maps. Finally, we aggregate activation maps with the same semantics and implement mitochondrial instance segmentation by a post-processing step.

3.2. Structural Template Learning Module

To make our model adapt to inputs of different appearances, our structural template learning module (STLM) can learn input-specific templates, which absorb and integrate information of background, foreground and contour. In specific, we introduce a set of initial templates with initial weight value satisfying the xavier distribution as $\mathbf{T} = \{\mathbf{T}^{\star}\} = \{\{t_i^{\star}\}_{i=1}^{\tau}\}$, where $\star \in (B, F, C)$ denotes the category of each template (background, foreground and contour), $t_i^{\star} \in \mathbb{R}^{1\times d}$ represents a classifier that determines whether voxels of the feature map belong to the i-th template of category $\star$, τ denotes the template number of each category.

For each layer in the structural template learning module, these templates, which are learnable parameters, are first feed into a self-attention layer, where all keys, queries and values arise from initial templates, to incorporate the local context of mitochondria. The updated templates then go through the cross-attention layer to extract specific semantics from the input voxel embedding, where queries arise from the templates, and keys and values arise from the input voxel embedding $\mathbf{f} \in \mathbb{R}^{h\times w\times d\times c}$, where h, w, d and c refer to the height, width, depth and channel number of the feature map, respectively. Formally,

$$\mathbf{Q}_n = \mathbf{T}\mathbf{W}^Q, \mathbf{K}_n = \mathbf{f}\mathbf{W}^K, \mathbf{V}_n = \mathbf{f}\mathbf{W}^V, \tag{1}$$

where $n \in [1, \ldots, N]$ and $\mathbf{W}^Q \in \mathbb{R}^{C\times C_k}$, $\mathbf{W}^K \in \mathbb{R}^{C\times C_k}$, $\mathbf{W}^V \in \mathbb{R}^{C\times C_v}$ are linear projections. The attention weights are calculated based on the dot-product similarity between each query and key:

$$m_{i,j}^{\star} = \frac{\exp\left(\beta_{i,j}^{\star}\right)}{\sum_{j=1}^{hwd}\exp\left(\beta_{i,j}^{\star}\right)}, \beta_{i,j}^{\star} = \frac{\mathbf{Q}_i^{\star}\mathbf{K}_j^T}{\sqrt{C_k}}, \tag{2}$$

where the attention weight $m_{i,j}^{\star}$ indicates the probability of the spatial feature belonging to the i-th template of the category $\star$. The attention weights of all hwd positions make up a type of prediction, which has high response values at voxels belonging to the corresponding template. We denote the activation maps of each template category $\star$ as:

$$M^{\star} = \begin{pmatrix} m_{1,1}^{\star} & m_{1,2}^{\star} & \cdots & m_{1,hwd}^{\star} \\ \vdots & & & \vdots \\ m_{\tau,1}^{\star} & m_{\tau,2}^{\star} & \cdots & m_{\tau,hwd}^{\star} \end{pmatrix}. \tag{3}$$

OT Regularization Term. Without additional constraints, the ambiguous activation in $M^{\star}$ will cause the confusion problem of different templates within the same category, perceiving repetitive area information. To alleviate this problem, as shown in the right of Figure 2, we introduce a regularization term to assign the group of semantically consistent voxels to the same template. Specifically, we formulate the template assignment problem as the optimal transport (OT) problem [48]. The goal of the OT problem is to find an optimal transportation plan $\mathbf{P}^*$ at a global minimal transportation cost, which can be solved in polynomial time

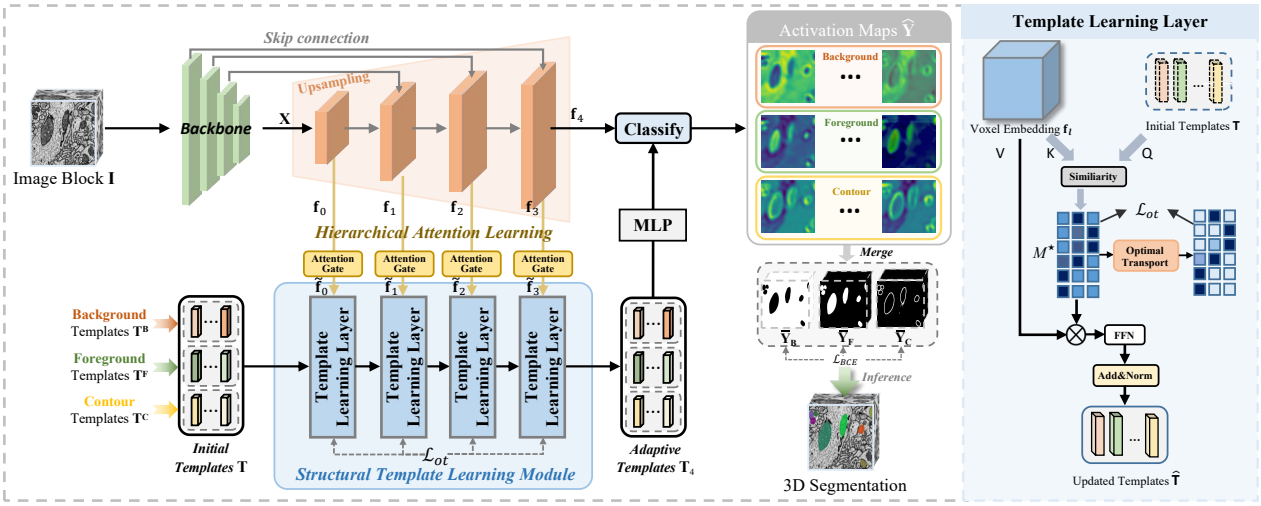


Figure 2. Framework of our proposed adaptive template transformer (ATFormer). It includes a structural template learning module (STLM) which learns a set of adaptive templates with specific semantic information, and a hierarchical attention learning mechanism (HALM) which aggregates multi-level voxel embeddings to adapt to mitochondria of all scales. The figure on the right illustrates the template learning layer in STLM.

by linear programming. In the template learning decoder, $M^\star \in \mathbb{R}^{\tau \times hwd}$ can serve as a preliminary assignment of each voxel. Then we can obtain the optimal assignment across all transportation plans by following the optimization function.

$$\mathbf{P}^* = \max_{\mathbf{P} \in \mathcal{P}} \operatorname{Tr}\left(\mathbf{P}^\top M^\star\right) + \epsilon H(\mathbf{P}), \tag{4}$$

where $H(\mathbf{P}) = -\sum_{ij} \mathbf{P}_{ij} \log \mathbf{P}_{ij}$ is the entropy function, and ϵ is the parameter that controls the smoothness of the assignment. We follow [1, 7] to enforce an prior partition μ by constraining the assignment matrix $\mathbf{P}$ to belong to the transportation polytope:

$$\mathcal{P} = \left\{ \mathbf{P} \in \mathbb{R}_+^{\tau \times hwd} \mid \mathbf{P}\mathbb{1} = \mu, \mathbf{P}^\top \mathbb{1} = \frac{1}{hwd} \cdot \mathbb{1} \right\}, \tag{5}$$

where $\mathbb{1}$ denotes the vector of all ones in the appropriate dimension. The optimal transport plan $\mathbf{P}^*$ can be obtained via a fast variant of Sinkhorn-Knopp [11]. $\mathbf{P}^*$ has the advantage of less ambiguous activations, which can further enlarge the discrepancy of diverse templates. We minimize the binary cross entropy loss between $\mathbf{P}^*$ and the output mask of the structural template learning decoder $M^\star$ to online refine foreground/background maps as following:

$$\begin{aligned} \mathcal{L}_{ot} = -\frac{1}{\tau hwd} \sum_{i=1}^{\tau} \sum_{j=1}^{hwd} & \mathbf{P}^*_{ij} \log M^\star_{ij} \\ & + \left(1 - \mathbf{P}^*_{ij}\right) \log\left(1 - M^\star_{ij}\right). \end{aligned} \tag{6}$$

Then, according to Eq. (2) and the values from the feature map, we can further obtain the updated templates:

$$t_i^\star = \mathrm{FFN}(\mathrm{Att}\left(\mathbf{Q}_i^\star, \mathbf{K}, \mathbf{V}\right)) = \mathrm{FFN}\left(\sum_{j=1}^{hwd} m_{i,j}^\star \mathbf{V}_j\right), \tag{7}$$

where it should be noted that each cross-attention layer and FFN is wrapped by a residual connection followed by layer normalization as in the transformer architecture.

3.3. Hierarchical Attention Learning Mechanism

To enable our learned adaptive templates to be scale-aware and aggregate long-range information, we further propose an efficient hierarchical strategy with attention gates to utilize multi-scale features of both low and high resolution to boost the activation ability of templates. Given the output embedding $\mathbf{f}_l$ from the upsampling module, where $l \in [0, \ldots, 3]$ (from the lowest to the highest level in order), the original feature is flattened, then pass through an attention gate before entering STLM. Each attention gate is composed of two consecutive layers including a multi-head self-attention (MSA) layer and a feed forward network (FFN) as follows:

$$\mathbf{f'}_l = \mathrm{MSA}(\mathrm{LN}(\mathbf{f}_l)) + \mathbf{f}_l, \tag{8}$$

$$\widetilde{\mathbf{f}}_l = \mathrm{FFN}(\mathrm{LN}(\mathbf{f'}_l)) + \mathbf{f'}_l, \tag{9}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization [3] following the transformer architecture.

Then, the modified feature map was transferred into the corresponding template learning layer in STLM as described in Sec. 3.2, to interact with template groups by

cross-attention. This operation will be repeated four times to produce the final adaptive templates:

$$\mathbf{T}_{l+1} = \mathrm{STLM}_{l+1}(\mathbf{T}_l, \widetilde{\mathbf{f}}_l), \tag{10}$$

where STLM_{l+1} denotes the corresponding layer in STLM. Please not that $\mathbf{f}_0$ refers to the bottom-level feature with the lowest resolution.

The obtained adaptive templates $\mathbf{T}_4 = \{\{\hat{t}_i^\star\}_{i=1}^{\tau}\}$ are used as classifiers to interact with the final voxel embedding $\mathbf{f}_4$ from the upsampling module, generating the activation maps of all three categories. The formula is as follows,

$$\hat{\mathbf{Y}} = \{\{\hat{y}_i^\star\}_{i=1}^{\tau}\}, \hat{y}_i^\star = \mathrm{MLP}(\hat{t}_i^\star)\mathbf{f}_4^\top, \tag{11}$$

where MLP comprises of two linear layers with GELU activation functions. For better use of the following training and inference, we need to merge the activation maps $\hat{\mathbf{Y}}$ into a probability map $\overline{\mathbf{Y}} \in \mathbb{R}^{3\times H\times W\times D}$ of background, foreground and contour. In our implementation, we treat the summation of background activation maps as background probability and do the same operation for foreground and contour, denoted as $\overline{\mathbf{Y}}_B$, $\overline{\mathbf{Y}}_F$ and $\overline{\mathbf{Y}}_C$.

3.4. Training and Inference

Loss Function. We follow the loss function used in [52]. The binary cross entropy (BCE) is a common loss function used in biomedical image segmentation. We impose supervisory constraints on background, foreground and contour predictions as:

$$\begin{aligned} \mathcal{L}_B &= \mathcal{L}_{BCE}(\overline{\mathbf{Y}}_B, \mathbf{Y}_B), \\ \mathcal{L}_F &= \mathcal{L}_{BCE}(\overline{\mathbf{Y}}_F, \mathbf{Y}_F), \\ \mathcal{L}_C &= \mathcal{L}_{BCE}(\overline{\mathbf{Y}}_C, \mathbf{Y}_C), \end{aligned} \tag{12}$$

where $\mathbf{Y}_B$, $\mathbf{Y}_F$ and $\mathbf{Y}_C$ are the corresponding ground-truth of $\overline{\mathbf{Y}}_B$, $\overline{\mathbf{Y}}_F$ and $\overline{\mathbf{Y}}_C$. The overall loss function L is defined as

$$\mathcal{L} = \mathcal{L}_B + \mathcal{L}_F + \mathcal{L}_C + \lambda_{ot}\mathcal{L}_{ot}, \tag{13}$$

where λ_{ot} denotes the coefficient of $\mathcal{L}_{ot}$.

Instance Inference. During inference, we obtain the predicted mask by applying an argmax operation on the probability map $\overline{\mathbf{Y}}$ without additional computational cost. It is noteworthy that this specific task requires the instance segmentation of mitochondria as a final result, in other words, treats each mitochondrion as a distinct individual instance, rather than solely segmenting all foreground regions as one entity. Therefore, we employ an efficient post-processing step following [22] to combine semantic predictions (background, foreground and contour) in the proposed model to generate the final instance segmentation.

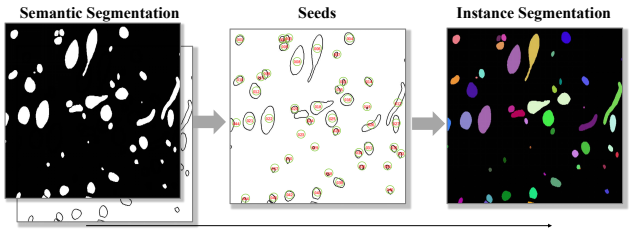


Figure 3. Illustration of instance inference. By using a post-processing step, the semantic predictions can be transformed into the final instance segmentation.

4. Experiments

4.1. Dataset

To demonstrate the effectiveness of our proposed model, we conduct extensive experiments on three benchmarks: MitoEM [52], Lucchi [30] and NucMM datasets [22].

MitoEM dataset is divided into two sub-datasets ($1000 \times 4096 \times 4096$ in voxels at $30 \times 8 \times 8$ nm resolution) dubbed as MitoEM-R and MitoEM-H, which contain mitochondria EM images of a rat and human tissue respectively. For the reason that the annotations of the last 500 slices are not open-sourced, we use the first 400 slices for training and evaluate the segmentation performance on the validation set containing the remaining 100 slices.

Lucchi dataset is a mitochondria semantic segmentation dataset providing segmentation results for the foreground. In our experiments, the training and testing data volumes are both with a size of $165 \times 1024 \times 768$.

NucMM-Z dataset is a nuclei dataset with 27 small chunks of size $64 \times 64 \times 64$ voxels from the zebrafish volume for training and another 27 volumes of the same size for testing. We use this dataset to evaluate the generalizability of our method.

4.2. Implementation Details

We adopt Pytorch [35] to implement the proposed method. 4 NVIDIA TITAN RTX (24GB) GPUs are used for training. For the MitoEM dataset, we use the architecture of 3D U-Net [10] as backbone with the input size of $32 \times 256 \times 256$ following [52]. During the training stage, our model is trained with the batch size of 8, using the Adam optimizer [27] with an initial learning rate of 0.0001 for 100,000 iterations. We set the number of templates in each category as $\tau = 3$. In the final loss function, we set $\lambda_{ot} = 0.5$. For the Lucchi dataset with only foreground ground-truth, we discard the use of the contour template, only train our model with foreground and background templates for the semantic mask output, following the training details in [52]. For the NucMM dataset, we directly input image volume into the model for the reason that all image volumes are isotropic voxels after sampling.

Method	MitoEM-R						MitoEM-H					
	mAP	AP50	AP75	APs	APm	APl	mAP	AP50	AP75	APs	APm	APl
U2D-B [52]	28.4	40.2	35.5	10.4	62.8	48.1	36.8	62.3	59.7	40.8	81.4	71.1
U3D-A [52]	26.5	38.4	32.8	40.8	23.5	65.3	42.1	65.5	61.7	56.4	77.4	61.7
U3D-BC [52]	45.6	57.3	52.1	29.0	75.1	49.0	45.5	66.2	60.5	48.9	82.0	61.8
Nightingale [33]	-	-	71.5	0.7	40.4	78.7	-	-	62.5	3.4	47.8	73.4
CLMS [21]	-	89.5	87.0	20.3	74.3	91.3	-	82.8	78.7	29.6	77.8	83.0
UNETR† [14]	70.3	89.7	83.1	17.3	81.7	87.3	61.4	83.1	73.9	23.8	73.2	75.8
Swin UNETR† [44]	73.7	95.0	90.4	30.1	84.3	93.4	62.0	87.9	80.3	57.7	83.1	79.2
ResUNet [20]	75.1	94.8	91.7	27.7	85.0	94.9	65.7	88.5	82.8	52.2	84.4	82.6
ATFormer(ours)	**78.2**	**96.2**	**92.8**	**38.7**	**85.8**	**96.3**	**68.2**	**89.7**	**84.1**	**57.9**	**85.2**	**83.9**

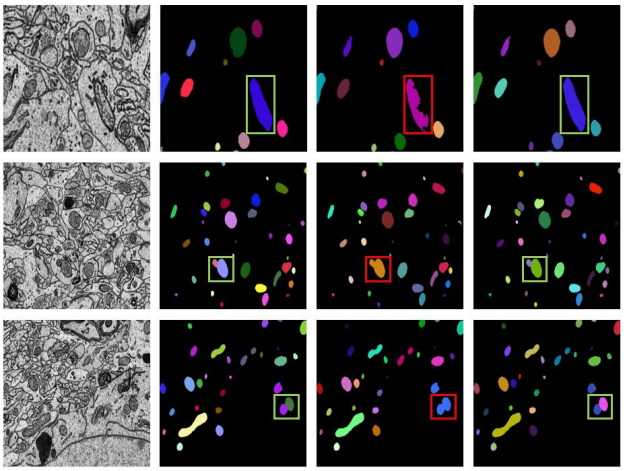
Table 1. Comparisons of different methods on the MitoEM-R and MitoEM-H [52] validation set, where † denotes the performance of our reproduction of the corresponding model.

Method	Jaccard	DSC
Lucchi [28]	75.5	86.0
Liu [23]	86.4	92.6
Yuan [57]	86.5	92.7
Wei [52]	88.7	-
Casser [8]	89.0	94.2
ResUNet [20]	89.5	94.5
ATFormer (Ours)	**90.2**	**94.8**

Table 2. Semantic segmentation results on the Lucchi [30] testing set.

Method	AP50	AP75
Cellpose [40]	79.6	34.2
StarDist [53]	91.2	32.8
U3D-BC [52]	78.2	55.6
U3D-BCD [22]	97.8	80.9
ATFormer (Ours)	**98.2**	**83.6**

Table 3. Instance segmentation results on the NucMM-Z [22] testing set.



(a) Image (b) Ground Truth (c) ResUNet (d) ATFormer (ours)

Figure 4. Qualitative comparison of different methods on the MitoEM-R and MitoEM-H validation set.

4.3. Evaluation Metrics

For a fair comparison, we adopt 3D mAP, AP50 and AP75 metric [52] on the MitoEM dataset. Please note that mAP is a more stringent metric than AP50 and AP75, for the reason that it takes into account the AP with iou greater than 75%. Besides, for better illustration of performance on mitochondria of different scales, we also show the results on APs, APm and APl which represents 3D AP75 of *small*, *medium* and *large* instances, respectively, divided by the volume threshold of 5K and 15K voxels. On the Lucchi dataset, we adopt metrics of jaccard-index coefficient (Jaccard) and dice similarity coefficient (DSC) to evaluate the effectiveness of semantic segmentation ability. For the NucMM-Z dataset, we use the same metrics as MitoEM dataset.

4.4. Main Results

4.4.1 Quantitative Evaluations

Our method outperforms the state-of-the-art methods for semantic and instance segmentation of mitochondria. For better demonstration, we reproduce the performance of SOTA methods in biomedical image analysis for the mitochondria task. From the retrained results of UNETR [14] and Swin UNETR [44], it can be observed that simply adopting the transformer architecture for feature extraction does not produce ideal predictions of mitochondria, which demonstrates the effectiveness of our custom design for this task. As shown in Table 1, on the MitoEM-R dataset, our method achieves a mean AP of 78.2%, AP50 of 96.2% and overall average AP75 of 92.8%, outperforming the second top-ranked methods by 3.1%, 1.4% and 1.1% respectively. With superior AP75 performance, it is noteworthy that the improvement on mAP is more significant, due to the fact that mAP is a more stringent evaluation metric that requires

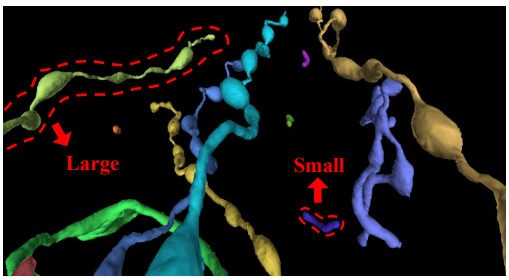


Figure 5. Visualization of reconstruction results from instance segmentation of our ATFormer on the MitoEM-R validation set.

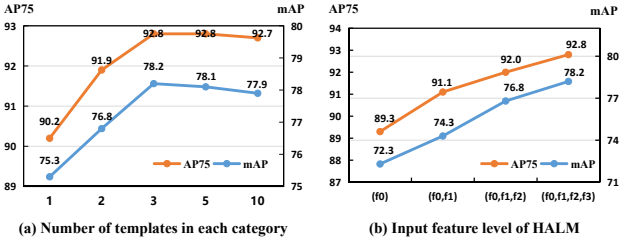


Figure 6. Comparisons of performance with different number of templates and input feature level of HALM on the MitoEM-R validation set in terms of mAP and AP75.

more complete segmentation results to be a true positive. This demonstrates the ability of our method to more fully activate the intact region based on the readiness to localize the mitochondrial location.

Besides, for mitochondria of different scales, our ATFormer demonstrates strong performance on all sizes of targets, thanks to the proposed adaptive scale-aware templates, improving performance on APs, APm and APl by 8.6%, 0.8% and 1.4%, respectively, over the second-place method on MitoEM-R. We can also observe that our method achieves a new SOTA performance with mAP of 68.2% and AP75 of 84.1% on the MitoEM-H validation set, with better recognition of mitochondria of all scales.

In Table 2, we further evaluate the effectiveness of ATFormer for the semantic segmentation task on the Lucchi dataset, with a clear performance gain on two metrics. As shown in Table 3, our proposed method performs favorably against benchmark results on the NucMM-Z dataset [22]. It can be observed that while existing methods can easily obtain the target location, segmenting the nucleus at a finer level is more challenging. Under these circumstances, our method is better able to fully activate the foreground with strong adaptability to various targets, achieving a performance improvement of 2.7% on AP75.

4.4.2 Qualitative Results

As shown in Figure 4, ATFormer shows improved segmentation performance for mitochondria under different circumstances. In specific, our method outperforms others in most scenarios, especially when the mitochondria occupy a relatively large space as shown in the 1^{st} row. ATFormer demonstrates a precise detection of foreground and contour against the surrounding background, which indicates the effectiveness of our proposed method. Besides, for two adjacent mitochondrial instances, ATFormer shows the excellent capability of separating them in such a situation as shown in the 2^{nd} and 3^{rd} row. Compared with existing models, ours exhibits higher boundary segmentation accuracy as it accurately identifies the boundaries between nearby instances. Besides, we demonstrate the results of dense instance segmentation of mitochondria in 3D space as shown in Figure 5. The results show that our ATFormer is able to distinguish adjacent instances and acquire ideal predictions.

4.5. Ablation Study

Can the adaptive templates improve model performance? Yes. The scale distribution of mitochondria can be hugely different, simply utilizing fixed classifiers for segmentation isn't sufficient enough for identifying some larger mitochondria. Thus in this case, adaptive scale-aware templates that adapt to the input can be more effective for this task. As shown in Table 4, with the utilization of adaptive templates, further improvements can be observed, *e.g.*, 5.2% in AP75, compared with the 1^{st} row and 3^{rd} row. During experiments, we notice that the number of templates in each category also affects the performance, as shown in Figure 6(a). It can be observed that our performance peaks at a template number of 3, however, continuing to increase the number causes a slight performance drop. We speculate that the reason is that as the number of templates increases, the OT regularization term keeps forcing each template within the same category to focus on a different region, which is contrary to the distribution of mitochondria, leading to model performance degradation.

Is the background information effective for mitochondria segmentation? Yes. It is undeniable that the foreground region is more significant for our desired segmentation goals, however, background information plays an integral role in highlighting mitochondria of interest in a reverse manner. From the 2^{nd} row and 3^{rd} row in Table 4, the introduction of the background template achieves a certain performance lift, which favorably manifests the foreground-background disambiguation effect. Activation maps in Figure 7(e) further prove our point.

How much does the OT regularization term contribute? As shown in Table 5, we construct some comparative experiments to explore the effectiveness of the OT regularization term. The removal of optimal transport means a naked structural template learning decoder without additional constraints. It can be seen that there is a drop in performance

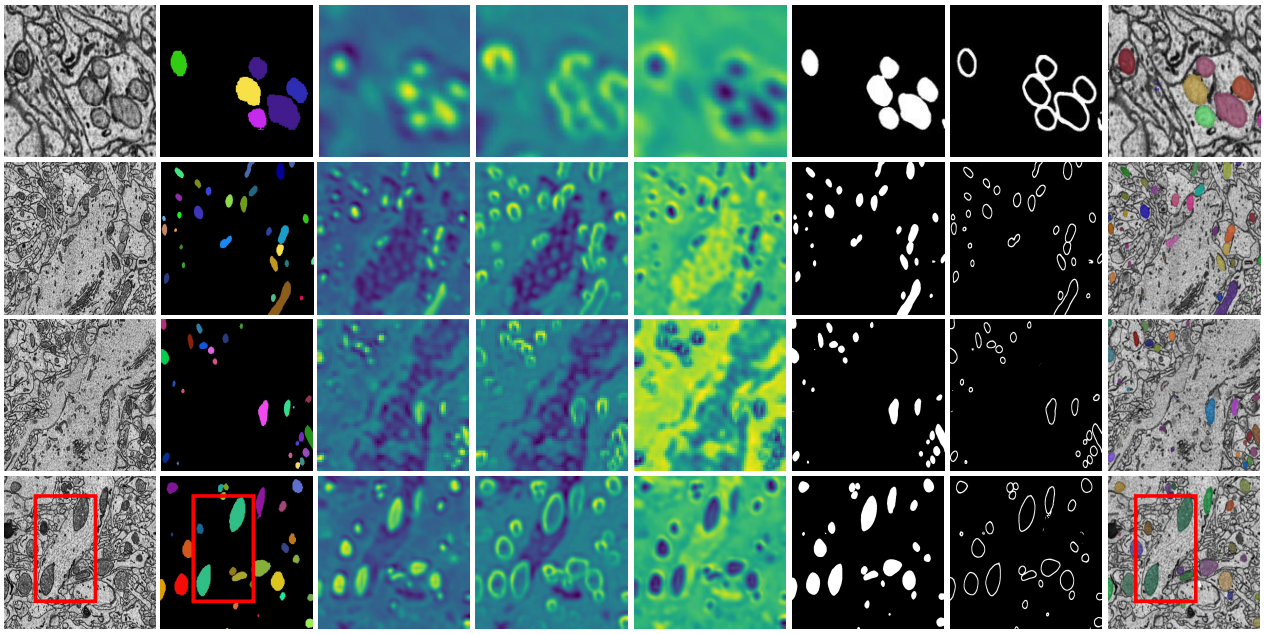


Figure 7. Visualization of segmentation results and activation maps of adaptive templates on the MitoEM-R validation set.

Templates			mAP	AP75
foreground	contour	background		
×	×	×	69.8	87.6
✓	✓	×	76.3	91.9
✓	✓	✓	**78.2**	**92.8**

Table 4. Ablation on different templates on the MitoEM-R validation set.

	mAP	AP75
ATFormer (ours)	**78.2**	**92.8**
−attention gate	76.6$_{(-1.6)}$	92.1$_{(-0.7)}$
−optimal transport	77.5$_{(-0.7)}$	92.4$_{(-0.4)}$
−both 2 components above	76.2$_{(-2.0)}$	91.8$_{(-1.0)}$

Table 5. Ablation on the OT regularization term and attention gates on the MitoEM-R validation set.

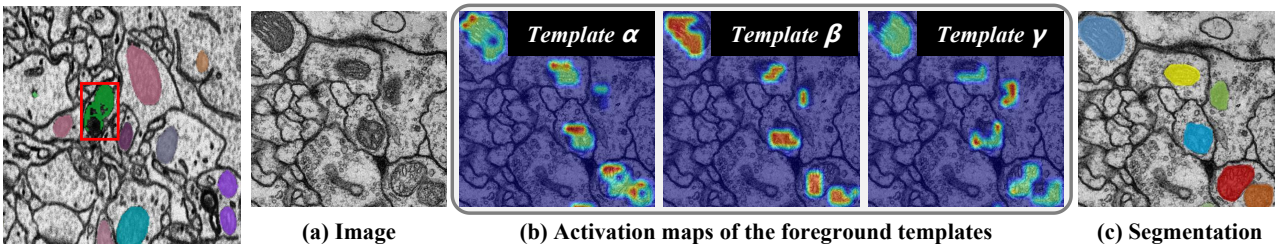
without the OT constraint, for the reason that the templates within the same category perceive information from repetitive regions, which are susceptible to the ambiguous activation of attention maps. Thus it is necessary to adopt the OT regularization term to enlarge the discrepancy among different templates.

Does the task of mitochondria segmentation require the hierarchical attention learning mechanism? Yes. The hierarchical attention learning mechanism aggregates multi-level features to enable our adaptive templates to be aware of scale variations. As shown in Figure 6(b), there are significant performance gains on both metrics when utilizing multi-level features compared to when using only the bottom embedding, demonstrating the effectiveness of HALM. Besides, we conduct ablation experiments on attention gates as shown in Table 5, illustrating that attention gates do bring a performance gain by self-attention mechanism to capture long-range contextual information. Together with the structural template learning module, HALM enables our model to generate adaptive scale-aware classifiers to segment various mitochondria at a finer level.

4.6. Explainable Visualization Study

We show more qualitative visualization results on the MitoEM dataset [52] in Figure 5 and 7, demonstrating robustness to mitochondria with large variations in appearance and scale. As shown in the activation maps of Figure 7, the adaptive templates successfully activate the region of the corresponding semantic category at a fine-grained level, producing accurate semantic and contour predictions, which significantly benefit the instance segmentation results. In addition to the strong ability of separating adjacent individuals apart as shown in the 1^{st} row, our ATFormer also demonstrates a significant long-range modeling capability. For example, as shown by the ground truth in the 4^{th} row, two bright green foregrounds in the red box are labeled as the same label. Although far apart in this slice, they do both belong to the same instance in 3D space due to the complex shape of mitochondria. Our ATFormer is able to determine that these two belong to the same instance accurately, thanks to the long-range modeling capability and strong adaptability of our method.

We further demonstrate the capability of adaptive tem-



(a) Image (b) Activation maps of the foreground templates (c) Segmentation

Figure 8. Failure case.

Figure 9. Visualization of activation maps with the aid of OT.

plates with activation maps of foreground for clarity as shown in Figure 9. With the help of the OT regularization term, the different templates within the same category focus on significant distinct areas. Besides, as shown in Figure 9, if the activation of a single template ($Template\ \beta$) is incomplete, other templates ($Template\ \alpha$ and γ) can be used to complement the activation, jointly producing intact semantic segmentation as shown in Figure 9 (c). This demonstrates the need for our multi-template design. Our method utilizes multiple adaptive templates to absorb contextual information, which is more robust to mitochondria of different sizes, thus yielding more precise segmentation.

Failure case and limitation analysis. Though achieving promising performance, our approach also struggles with challenging scenarios (*e.g.*, over-adjacent instances, and unexpected stains during imaging) as shown in Figure 8. We believe that the patterns of these failure cases also shed light on the possible direction of our future research. The overall methods in the field are non-end-to-end, which generally require post-processing steps to produce instances. End-to-end instance segmentation without post-processing remains a direction worth exploring.

5. Conclusion

In this paper, we introduce a novel adaptive template transformer (ATFormer) for mitochondria segmentation in EM images. Specifically, we design a structural template learning module to acquire appearance-adaptive templates and a hierarchical attention learning mechanism to aggregate multi-scale information for adapting to mitochondria with large variations in appearance and scale. Besides, an OT regularization term is proposed to enlarge the discrepancy among diverse templates for full activation. Extensive experimental results demonstrate the effectiveness of our proposed ATFormer for mitochondria segmentation.

Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (Grant 62022078, 62121002) and the National Defense Basic Scientific Research Program (Grant JCKY2021130B016).

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019. 4

[2] Reza Azad, Mohammad T Al-Antary, Moein Heidari, and Dorit Merhof. Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. *IEEE Access*, 10:108205–108215, 2022. 2

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] Ella Bossy-Wetzel, Mark J Barsoum, Adam Godzik, Robert Schwarzenbacher, and Stuart A Lipton. Mitochondrial fission in apoptosis, neurodegeneration and aging. *Current opinion in cell biology*, 15(6):706–716, 2003. 1

[5] Ana Bratic, Nils-Göran Larsson, et al. The role of mitochondria in aging. *The Journal of clinical investigation*, 123(3):951–957, 2013. 1

[6] Silvia Campello and Luca Scorrano. Mitochondrial shape changes: orchestrating cell pathophysiology. *EMBO reports*, 11(9):678–684, 2010. 1

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 4

[8] Vincent Casser, Kai Kang, Hanspeter Pfister, and Daniel Haehn. Fast mitochondria detection for connectomics. In *Medical Imaging with Deep Learning*, pages 111–120. PMLR, 2020. 2, 6

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[10] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2, 5

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013. 4

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[13] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pages 272–284. Springer, 2022. 2

[14] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022. 2, 3, 6

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[16] Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6202–6212, 2023. 2

[17] Wei Huang, Xiaoyu Liu, Zhen Cheng, Yueyi Zhang, and Zhiwei Xiong. Domain adaptive mitochondria segmentation via enforcing inter-section consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 89–98. Springer, 2022. 2

[18] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020. 3

[19] Andrew B Knott, Guy Perkins, Robert Schwarzenbacher, and Ella Bossy-Wetzel. Mitochondrial fragmentation in neurodegeneration. *Nature Reviews Neuroscience*, 9(7):505–518, 2008. 1

[20] Mingxing Li, Chang Chen, Xiaoyu Liu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Advanced deep networks for 3d mitochondria instance segmentation. *arXiv preprint arXiv:2104.07961*, 2021. 2, 6

[21] Zhili Li, Xuejin Chen, Jie Zhao, and Zhiwei Xiong. Contrastive learning for mitochondria segmentation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3496–3500. IEEE, 2021. 6

[22] Zudi Lin, Donglai Wei, Mariela D Petkova, Yuelong Wu, Zergham Ahmed, Silin Zou, Nils Wendt, Jonathan Boulanger-Weill, Xueying Wang, Nagaraju Dhanyasi, et al. Nucmm dataset: 3d neuronal nuclei instance segmentation at sub-cubic millimeter scale. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 164–174. Springer, 2021. 5, 6, 7

[23] Jing Liu, Linlin Li, Yang Yang, Bei Hong, Xi Chen, Qiwei Xie, and Hua Han. Automatic reconstruction of mitochondria and endoplasmic reticulum in electron microscopy volumes by deep learning. *Frontiers in neuroscience*, 14:599, 2020. 6

[24] Jing Liu, Weifu Li, Chi Xiao, Bei Hong, Qiwei Xie, and Hua Han. Automatic detection and segmentation of mitochondria from sem images using deep neural network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 628–631. IEEE, 2018. 2

[25] Xiaoyu Liu, Bo Hu, Mingxing Li, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. A soma segmentation benchmark in full adult fly brain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7402–7411, 2023. 3

[26] Xiaoyu Liu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Biological instance segmentation with a superpixel-guided graph. IJCAI, 2022. 3

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[28] Aurélien Lucchi, Yunpeng Li, and Pascal Fua. Learning for structured prediction using approximate subgradient descent with working sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1987–1994, 2013. 6

[29] Aurélien Lucchi, Pablo Márquez-Neila, Carlos Becker, Yunpeng Li, Kevin Smith, Graham Knott, and Pascal Fua. Learning structured models for segmentation of 2-d and 3-d imagery. *IEEE transactions on medical imaging*, 34(5):1096–1110, 2014. 2

[30] Aurélien Lucchi, Kevin Smith, Radhakrishna Achanta, Graham Knott, and Pascal Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE transactions on medical imaging*, 31(2):474–486, 2011. 5, 6

[31] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17927, 2023. 3

[32] Huayu Mai, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Dualrel: Semi-supervised mitochondria segmentation from a prototype perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19617–19626, 2023. 2

[33] Luke Nightingale, Joost de Folter, Helen Spiers, Amy Strange, Lucy M Collinson, and Martin L Jones. Automatic instance segmentation of mitochondria in electron microscopy data. *bioRxiv*, 2021. 6

[34] Ismail Oztel, Gozde Yolcu, Ilker Ersoy, Tommi White, and Filiz Bunyak. Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1195–1200. IEEE, 2017. 2

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[36] Martin Picard, Douglas C Wallace, and Yan Burelle. The rise of mitochondria in medicine. *Mitochondrion*, 30:105–116, 2016. 1

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[38] Richard Schalek, Dongil Lee, Narayanan Kasthuri, Adi Peleg, T Jones, Verena Kaynig, Daniel Haehn, Hanspeter Pfister, David Cox, and Jeff W Lichtman. Imaging a 1 mm 3 volume of rat cortex using a multibeam sem. *Microscopy and Microanalysis*, 22(S3):582–583, 2016. 1

[39] Mojtaba Seyedhosseini, Mark H Ellisman, and Tolga Tasdizen. Segmentation of mitochondria in electron microscopy images using algebraic curves. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 860–863. IEEE, 2013. 2

[40] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021. 2, 6

[41] Adi Suissa-Peleg, Daniel Haehn, Seymour Knowles-Barley, Verena Kaynig, Thouis R Jones, Alyssa Wilson, Richard Schalek, Jeffery W Lichtman, and Hanspeter Pfister. Automatic neural reconstruction from petavoxel of electron microscopy data. *Microscopy and Microanalysis*, 22(S3):536–537, 2016. 1

[42] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021. 3

[43] Rui Sun, Naisong Luo, Yuwen Pan, Huayu Mai, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Appearance prompt vision transformer for connectome reconstruction. IJCAI, 2023. 3

[44] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022. 3, 6

[45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[47] Amelio Vazquez-Reina, Michael Gelbart, Daniel Huang, Jeff Lichtman, Eric Miller, and Hanspeter Pfister. Segmentation fusion for connectomics. In *2011 International Conference on Computer Vision*, pages 177–184. IEEE, 2011. 2

[48] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. 3

[49] Douglas C Wallace. Mitochondria and cancer. *Nature Reviews Cancer*, 12(10):685–698, 2012. 1

[50] Yuan Wang, Rui Sun, and Tianzhu Zhang. Rethinking the correlation in few-shot segmentation: A buoys view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2023. 3

[51] Li Wangkai, Li Zhaoyang, Sun Rui, Mai Huayu, Luo Naisong, Yuan Wang, Pan Yuwen, Xiong Guoxin, Lai Huakai, Xiong Zhiwei, et al. Maunet: Modality-aware anti-ambiguity u-net for multi-modality cell segmentation. In *Competitions in Neural Information Processing Systems*, pages 1–12. PMLR, 2023. 3

[52] Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, et al. Mitoem dataset: large-scale 3d mitochondria instance segmentation from em images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–76. Springer, 2020. 1, 2, 5, 6, 8

[53] Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, and Gene Myers. Star-convex polyhedra for 3d object detection and segmentation in microscopy. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3666–3673, 2020. 6

[54] Chi Xiao, Xi Chen, Weifu Li, Linlin Li, Lu Wang, Qiwei Xie, and Hua Han. Automatic mitochondria segmentation for em data using a 3d supervised convolutional network. *Frontiers in neuroanatomy*, 12:92, 2018. 2

[55] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3

[56] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 3

[57] Zhimin Yuan, Jiajin Yi, Zhengrong Luo, Zhongdao Jia, and Jialin Peng. Em-net: Centerline-aware mitochondria segmentation in em images via hierarchical view-ensemble convolutional network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1219–1222. IEEE, 2020. 6